UDC: 004.93 DOI: https://doi.org/10.15407/jai2021.02.08

DATA PROCESSING ALGORITHMS FOR THE IN SILICO SARS-COV-2 EPITOPE PREDICTION AND VACCINE DEVELOPMENT

M. Sprindzuk[1], **A. Vladyko**[2], **L. Titov**[3]
[1] Belarussian State University of Informatics and Radioelectronics, Belarus
4, P. Brovki st., Minsk, 220013
The State Scientific Institution «The United Insitute of Informatics Problems
of the National Academy of Sciences of Belarus», Belarus
6, Surganova st., Minsk, 220012
[2,3] The Republican Research and Practical Center for Epidemiology and Microbiology, Belarus
23, Filimonova st., Minsk, 220114

[1]https://orcid.org/0000-0001-9500-2954
[2]https://orcid.org/0000-0001-6927-5043
[3]https://orcid.org/0000-0002-4524-3730

This work was supported by the Belarusian Republican Foundation for Fundamental Research in the framework of the project*:*
M21COVID-001 "Development and screening of anti-COVID-19 mucosal vaccine based on the vector platform of intestinal adenovirus",
No. GR 20210889 dated 04.26.2021

Abstract. Based on literature analysis and own bioinformatics and virology research experience, authors propose multistep data processing algorithms, designed for the objectives of assisting the SARS-CoV-2 epitope vaccine production.

Epitope vaccines are expected to provoke a weaker but safer response of the vaccinated person. Methodologies of reverse bioengineering, vaccinology and synthetic peptide manufacturing have a promising future to combat COVID-19 brutal disease.

The significant mutational variability and evolution of the SARS-CoV-2, which is more typical for natural animal-borne viruses, are the hurdle for the effective and robust vaccine application and therefore require multidisciplinary research and prevention measures on the international level of cooperation.

However, we can expect that other viruses with different nature and content may be labelled as SARS-CoV-2. In this case metagenomics is an important discipline for COVID-19 discovery.

High quality reliable virus detection is still an unresolved question for improvement and optimization.

It is of upmost importance to develop the in silico and in vitro methods for the vaccine recipient reaction prediction and monitoring as techniques of the so-called modern personalized medicine.

Many questions can`t be solved applying exclusively in silico techniques and only can be discovered in vitro and in vivo, demanding significant time and money investments.

Future experiments also should be directed at the discovery of optimal vaccine adjuvants, vectors and epitope ensembles, as well as the personal characteristics of citizens of a certain region. This research would require several more years of meticulous large-scale laboratory and clinical work in various centers of biomedical institutions worldwide.

Keywords: SARS-CoV-2, COVID-19, epitope vaccine, medical cybernetics, bioinformatics, genomics, algorithms.

АЛГОРИТМИ ОБРОБКИ ДАНИХ ДЛЯ ПРОГНОЗУВАННЯ ЕПІТОПУ SARS-COV-2 МЕТОДОМ IN SILICO ТА РОЗРОБКИ ВАКЦИНИ

М. В. Спринджук[1], **О. С. Владико**[2], **Л. П. Титов**[3]
[1] Білоруський державний університет інформатики та радіоелектроніки, Білорусь
вул. П. Бровки, 4, м.Мінськ, 220013
Державна наукова установа «Об'єднаний інститут проблем інформатики
Національної академії наук Білорусі», Білорусь
вул. Сурганова, 6, м.Мінськ, 220012

[2,3] Республіканський науково-практичний центр епідеміології і мікробіології, Білорусь
вул. Філімонова, 23, м.Мінськ, 220114

[1]https://orcid.org/0000-0001-9500-2954
[2]https://orcid.org/0000-0001-6927-5043
[3]https://orcid.org/0000-0002-4524-3730

Робота виконана за підтримки Білоруського республіканського фонду фундаментальних досліджень у рамках проєкту:
M21COVID-001 «Розробка та скринінг мукозної вакцини проти COVID-19 на основі векторної платформи кишкового аденовірусу», № ГР 20210889 від 26.04.2021

Анотація. На основі аналізу літературних джерел та власного досвіду досліджень у галузі біоінформатики та вірусології, автори пропонують багатоетапні алгоритми обробки даних, що розроблені з метою сприяння виробництву епітопної вакцини проти SARS-CoV-2.

Очікується, що епітопні вакцини будуть безпечнішими й викликатимуть більш слабку реакцію організму. Методології зворотної біоінженерії, вакцинології та виробництва синтетичних пептидів мають велике майбутнє у боротьбі з тяжким захворюванням COVID-19.

Значна генетична мінливість та еволюція вірусу SARS-CoV-2, що притаманна природним вірусам, які походять від тварин, є перешкодою для ефективного та надійного застосування вакцини, а тому вимагає мультидисциплінарних досліджень та профілактичних заходів на міжнародному рівні співпраці.

Однак інші віруси з відмінною природою та будовою можуть бути позначені як SARS-CoV-2. Отже, метагеноміка є важливою дисципліною для виявлення COVID-19.

Питання якісного та надійного виявлення вірусів залишається відкритим для вдосконалення та оптимізації.

Надзвичайно важливою є розробка методів in silico та in vitro для прогнозування та моніторингу реакції реципієнта вакцини, як методик так званої сучасної персоналізованої медицини.

Застосування виключно методу in silico недостатньо для вирішення багатьох питань. Ці проблеми потребують застосування методів in vitro та in vivo, що вимагають значних затрат часу та коштів.
Майбутні експерименти також мають бути спрямовані на виявлення оптимальних ад’ювантів, векторів та поєднань епітопів, а також на індивідуальні особливості мешканців певного регіону. Для цього дослідження знадобиться ще кілька років ретельної й масштабної лабораторної та клінічної роботи в різних центрах біомедичних установ по всьому світу

Ключові слова: SARS-CoV-2, COVID-19, епітопна вакцина, медична кібернетика, біоінформатика, геноміка, алгоритми.

Introduction, problem and objectives review

The development of algorithms for processing data from genomes of especially dangerous viruses has already been partially covered in scientific publications by authors from the largest research centers of virology and tropical infections.

Several sources present algorithms for the analysis of genomic texts of the Zika virus, West Nile fever, Chikungunya, Nipah, Ebola, African swine fever, MERS (Middle East respiratory syndrome), SARS (severe acute respiratory syndrome), preceding the current COVID-19 pandemic (Coronavirus disease 2019).

Over the past six months, a large number of articles and chapters of monographs on the topic of algorithms for processing genomic data for the goals of development of a rational design of anti SARS-CoV-2 peptide vaccine have been appeared in print [1-13].

SARS-CoV-2 is a modern plague, which has been killing people worldwide since 2019. The estimated number of related deaths is about 5 000 000.

The exact origin of the virus is still unknown, several hypotheses propose deforestation and zoonoses together with recombinant and artificial virus leakages as a possible cause of the pandemics.

Thus, the development of effective and safe vaccines, chemoprohylaxis and therapy is of upmost importance today.

Methodology

Reverse vaccinology allows to use genome texts in order to define and select the essential genome regions, applying the modern bioinformatics and genomics software and algorithms, for the final goals of synthetic peptide bioengineering, vaccine design production, testing and practical application with an expected protective effect and benefits for the consumers.

Viral genome texts are produced via the signal processing of the sequencing machines (Illumina, Ion Torrent, Oxford Nanopore MinIon, Pacific Biosciences, etc.) and the resultant raw data are subjected to the multistep bioinformatics data processing pipeline.

The final result of the data processing are the genome contigs with an annotation, assigning the biologically and pathogenetically relevant terms and definitions to the input chunks of nucleotide or the amino acid words.

Therefore, we have to understand that the genomics data processing way is an error prone activity, which is highly dependent on the input data quality, proper de novo assembly algorithms and their options and settings and the relevant contig annotations.

Results and discussion

Based on the 13 years of experience of biology and relevant bioinformatics computations and the extensive modern literature on the topics of the reverse vaccinology, next generation sequencing data processing, we have developed data processing algorithms for the objectives of the SARS-CoV-2 epitope compound design (Figures 1-4).

The proposed algorithms may be implemented applying various software and it is realistic to expect the variations of the resultant epitope amino acid sequences therefore.

Figure 4 shows how the genome contigs can be annotated to disclose their physiological and biological functions.

When the elements of genome texts, representing the highly immunogenic and simultaneously non-allergenic to the host epitopes are defined, they are further processed incorporating solid-state peptide synthesis for producing vaccine components.

However, the generated synthetic proteins are not the complete vaccine preparation and is just an artificial analogue of the actual virus.

Further experiments are required and expectantly will show and substantiate the following questions, regarding the successful compositions of the multiepitope vaccine:

1. How the synthetic peptides affect the immune response of the mammals. Including the human host organism, how it compares versus the natural response of the actual SARS-CoV-2 virus and the other vaccines?
2. How the application of the few-epitope and multiepitope vaccines compares and may the epitope overload provoke the excessive immune reactions, such as the previously described cytokine storm, anaphylaxis and degenerating autoimmune reactions?
3. How the vaccine adjuvants and vectors change the immune response and the disease prevention efficacy in vitro and in vivo?

These are only a few questions to be replied urgently.

The quality of the initial data is of decisive importance for in silico bioinformatics analysis of genomes. Errors in reference databases can lead to incorrect amino acid substitutions, which will result in the synthesis of an ineffective immunogene that is unable to bind antigens and produce a sufficiently strong antiviral immune response.

Conclusion

On the basis of scientific literature data and the experience of computational experiments, algorithms for processing coronavirus genomes have been developed for the purposes and tasks of modern immunoinformatics, vaccinomics and virology. The algorithms can be applied and adapted to develop epitope vaccines against highly dangerous viruses of various origins. To implement the developed algorithms, various software and its elements, ensembles and complexes can be used. Viral vaccine development is a time and money highly

demanding research and manufacturing endeavor and requires serious quality control measures and the unexpected event and complications registration and reporting. The benefits of vaccination is not guaranteed for the whole population of consumers.

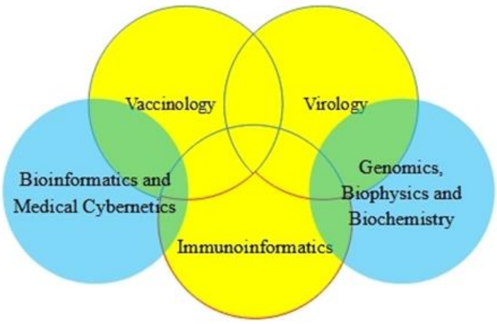

Figure 1. Venn diagram showing interdisciplinary nature of the SARS-CoV-2 epitope vaccine development

The major role of scientific efforts is to assist the development of safe vaccines to reduce the number of vaccine-related deaths and severe health harms.

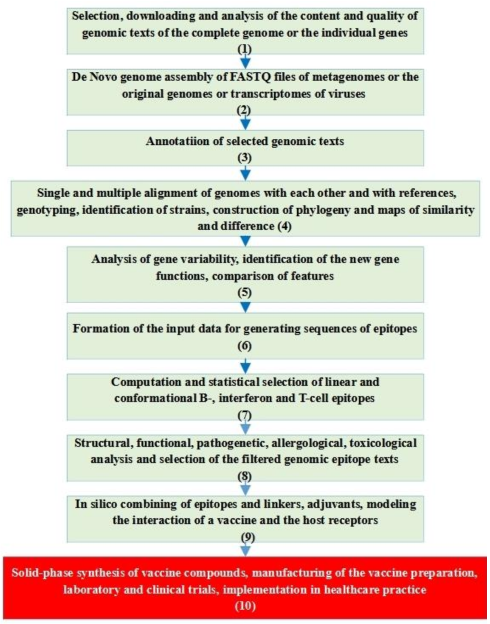

Figure 2. Technical presentation of the principal algorithm for the development of an epitope vaccine

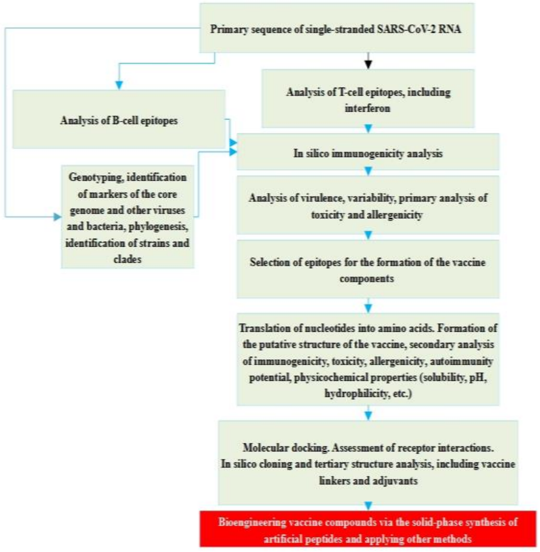

Figure 3. Biomedical presentation of the general algorithm for the development of an epitope vaccine

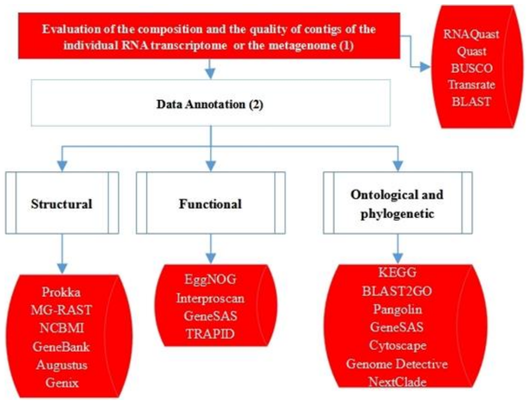

Figure 4. Genome annotation classification and software

M.V. Sprindzuk, A.S. Vladyko, L.P. Titov
Data processing algorithms for the in silico SARS-CoV-2 epitope prediction and vaccine development

Based on literature analysis and own bioinformatics and virology research experience, authors propose multistep data processing algorithms, designed for the objectives of assisting the SARS-CoV-2 epitope vaccine production. Many questions can`t be solved applying exclusively in silico techniques and only can be discovered in vitro and in vivo, demanding significant time and money investments.

List of terms and definitions and terms (with additions and changes from) [14]:

- antigenicity – the property to induce an immune response;
- antigens – inducers of the immune response, modifiers of the body's immunological reactivity;
- genome – a complete set of genes that determine all the properties of an organism;
- genetic recombination – the process of formation of genomes containing genetic material from two or more parental forms;
- genotype – a set of active and inactivated genes that are the principal part of the factors of heredity of the organism;
- genetic engineering – a branch of genetics that develops techniques for manipulating nucleic acids and uses these methods for genetic research and obtaining organisms with mixed genomes;
- gene pool – a set of genotypes of individuals representing a population, a type of microorganism and other organic forms;
- genophore – a carrier of genes;
- identification of viruses – laboratory and bioinformatics process of determining the systematic position of an unknown virus strain down to the type or variant;
- isolates – cultures of viruses or other microbes isolated from a specific source;
- immunization – a way to artificially create immunity;
- immunity – a complex of protective and adaptive reactions and adaptations aimed at maintaining the constancy of the antigenic composition of the internal environment of the macroorganism by killing and other types of neutralization and removal of foreign objects of antigenic nature;
- humoral immunity – immunity, the major effectors of which is antibodies;
- cellular immunity – immunity, the principal effector of which are sensitized lymphocytes and the lymphokines produced by them;
- acquired immunity – a form of immunity that is acquired in the process of individual development of the organism as a result of contact with parasites and substances of antigenic nature;
- interferons – a class of inductive low molecular weight alpha-helical proteins of vertebrates, possessing antiviral and other activities within the species to which the producer of interferons belongs;
- infection – a set of pathological, adaptive and reparative reactions of an organism resulting from its competitive interaction with microbes;
- pathogenicity of viruses – the species potential ability of viruses to cause an infectious process in their hosts;
- reinfection – recurrent infection of an organism that has suffered a disease with the same or another variant of a certain type of pathogen;
- superinfection – infection of a patient with the same or another variant of the same pathogen during the course of the disease;
- transcription – the process of transferring genetic information from the genome to informational RNA;
- translation – the process of formation of a polypeptide chain on mRNA associated with polyribosomes;
- transfection – infection of cells by introducing genomic and subgenomic molecules of nucleid acids.

References

1. Frishman, D., Marz, M. Virus Bioinformatics.–: CRC Press, 2021.– 296. 1000426548.
2. Ezzemani, W., Windisch, M.P., Kettani, A. et al. Immuno-informatics-based identification of novel potential B cell and T cell epitopes to fight Zika virus infections // Infect Disord Drug Targets. – 2020. – Vol 8, Online ahead of print, № 1. – P. 10. IDDT-EPUB-109015
[pii] 10.2174/1871526520666200810153657.
3. Alam, A., Ali, S., Ahamad, S. et al. From ZikV genome to vaccine: in silico approach for the epitope-based peptide vaccine against Zika virus envelope glycoprotein // Immunology. – 2016. – Vol 149, № 4. – P. 386-399. 10.1111/imm.12656.
4. Janahi, E.M., Dhasmana, A., Srivastava, V. et al. In silico CD4+, CD8+ T-cell and B-cell immunity associated immunogenic epitope prediction and HLA distribution analysis of Zika virus // EXCLI J. – 2017.–Vol 16,–P. 63-72.10.17179/excli2016-719.
5. Waller, F.M., Reche, P.A., Flower, D.R. West Nile Virus Vaccine Design by T Cell Epitope Selection: In Silico Analysis of Conservation, Functional Cross-Reactivity with the Human Genome, and Population Coverage // J Immunol Res. – 2020. – Vol 2020, – P. 7235742. 10.1155/2020/7235742.

6. Dutta, S.K., Bhattacharya, T., Tripathi, A. Chikungunya virus: genomic microevolution in Eastern India and its in-silico epitope prediction // 3 Biotech. – Vol 8, № 7. – P. 318. 10.1007/s13205-018-1339-3.
7. Pratheek, B.M., Suryawanshi, A.R., Chattopadhyay, S. In silico analysis of MHC-I restricted epitopes of Chikungunya virus proteins: Implication in understanding anti-CHIKV CD8(+) T cell response and advancement of epitope based immunotherapy for CHIKV infection // Infect Genet Evol.–2015.–Vol 31.– P118-26.10.1016/j.meegid.2015.01.017.
8. Sakib, M.S., Islam, M.R., Hasan, A.K., Nabi, A.H. Prediction of epitope-based peptides for the utility of vaccine development from fusion and glycoprotein of nipah virus using in silico approach // Adv Bioinformatics. – 2014. – Vol 2014, – P. 402492. 10.1155/2014/402492.
9. Majee, P., Jain, N., Kumar, A. Designing of a multi-epitope vaccine candidate against Nipah virus by in silico approach: a putative prophylactic solution for the deadly virus // J Biomol Struct Dyn. – 2021. – Vol 39, № 4. – P.1461-1480. 10.1080/07391102.2020.1734088.
10. Ali, M.T., Islam, M.O. A Highly Conserved GEQYQQLR Epitope Has Been Identified in the Nucleoprotein of Ebola Virus by Using an In Silico Approach // Adv Bioinformatics. – 2015. – Vol 2015. – P. 278197. 10.1155/2015/278197.
11. Mima, K.A., Katorkina, E., Katorkin, S. et al. In silico identifikacija B-i T-kletochnyh jepitopov belka CD2v virusa afrikanskoj chumy svinej (African swine fever virus, Asfivirus, Asfarviridae) // Voprosy virusologii. – 2020. – Vol 65, № 2. – P. 103-112.
12. Mahmud, S., Rafi, M. O., Paul, G. K. et al. Designing a multi-epitope vaccine candidate to combat MERS-CoV by employing an immunoinformatics approach // Sci Rep.–2021.Vol 11, № 1.–P.15431.10.1038/s41598-021-92176-1.
13. Sanami, S., Zandi, M., Pourhossein, B. et al. Design of a multi-epitope vaccine against SARS-CoV-2 using immunoinformatics approach // Int J Biol Macromol. – 2020. – Vol 164, – P. 871-883. S0141-8130(20)33870-8 [pii].
14. Krasilnikov, A.P. Mikrobiologicheskij slovar-spravochnik. - 2-e izd., dop. i pererab. - Mn. : OOO "Asar", 1999. - 397 C. 985-6070-50-3.

Received 25.11.2021
Accepted 30.11.2021